



Team H

Abhishek Jain

Bo Wang

Rashmi Upadhyaya

Yuyue Jiang

**Executive Summary**

Energy saving is always a big concern to every country. In United States, not only consumers who want to save up on expenses associated with utilities, but government and NGO care about protecting the earth by cutting down the energy consumption. When it comes to electricity saving, people often fail to understand what parameters to control to save electricity.

This analysis examines the relationship between electricity usage, residential building attributes and demographic characteristics using the U.S. Department of Energy’s Residential Energy Consumption Survey taken in 1995. We applied both parametric and nonparametric models to classify residents into ‘high’ electricity consumers or ‘low’ electricity consumers and also to predict which factors of a household have maximum effect on the consumption of electricity.

The default model, which proves to be the worst case scenario does not accommodate use of any such features. Our study compares different models suggested against the default model for both classification and prediction of electricity consumption and shows the amount of improvement for each model. From them, we select the models which provides us the highest accuracy and minimum error for our task.

We evaluate the impact of 50 features of household on electricity consumption. These features include type of housing unit, area of house, people staying in house, geographic location, etc. Our study relies on the data set provided and results have been derived assuming data set has captured correct values as answered by the user .

In order to achieve more effectiveness and lessen the chances of overfitting, we used randomized samples of data for both training and testing our models using 10-fold cross validation. This study compares various models used after feature selection (using PCA and Random Forest) on the basis of root mean squared error in case of prediction. We used various classifiers, to categorize the energy consumption into high/low energy, to segregate both the groups and analyzed true positive, false positive and ROC characteristics for all the classifiers.

The result of this analysis would help in prediction of approximate electric energy consumption on the basis of various household demographic parameters and also distinguish resident of house to which category of high or low energy consumers. In addition, this study has potential application in cost-benefit analysis of possible residential upgrades or purchases, and which other factors need change in order to cut down on electricity bills. Finally this analysis can really be helpful for the customers who are willing to buy a new house, and any upcoming society projects.

**Project Goals Tasks and Values**

To help saving residential electricity consumption, we have taken a deep analysis into the data from U.S. Department of Energy’s Residential Energy Consumption Survey taken in 1995. Using intuition, knowledge and professional data mining knowledge, we plan to execute prediction and classification of electricity consumption, and get most important features related. Based on the reality and final result, we give out the business value which contributes to electricity saving.

**1. Data Preprocessing**

We were given three datasets: Energy Consumption data (32 variables), Household Characteristics data (29 variables), and Housing Unit data (30 variables) with 4382 observations. The information on each dataset is as follows :

1. Energy consumption - records of the annual household consumption of various energies including electricity, natural gas, fuel oil.
2. Household data - demographics information about the people who live in the house including the employment, income, age.
3. Housing table - data about the wall materials, type of neighborhood the house is located in, the total area of the house and the area of separate parts of the house.

For our study, we merged the three tables by the identification number. We only focus on predicting the total electricity consumption since it is the dominant source of the total energies. Some of the variables, we noted had features with more than 95% inapplicable values and were not affecting the consumption and we decided to exclude them. We identified one row having incorrect values for the total square foot of the housing unit so we removed it.

So the initial dataset of 92 variables and 4382 rows is now converted into a dataset with 51 variables and 4381 rows with one being the output variable “BTUEL” (Electric annual use In thousands of BTU), and 50 being the input features.

**2. Prediction Analysis**

Predictive analysis exploits various patterns found in historical data to identify relationships among many factors, which could guide effective decision making. To accomplish this task, we analyzed various techniques of feature selection and reduction and predicted energy consumption using different models.

**2.1 Feature Selection and Reduction**

In order to avoid overfitting and express most of the variance of the dataset, we first implement feature selection and reduction.

**2.1.1 Random Forest**

Random forest also known as random decision forest operate by constructing multiple decision trees to overcome decision tree’s habit of overfitting on training data.

We started with training the random forest model varying various parameters which could possibly be the best fit for our data. We varied number of trees(ntree) along with the depth of the tree using the combination of number of variables randomly sampled as candidates at each split (mtry), minimum size of terminal nodes(nodesize) and maximum number of terminal nodes trees in the forest can have(maxnodes). After looping between optimal ranges and trying out various combinations and trying out different combination of values we concluded with values that gave us least RMSE and maximum R-Squared values running random forest regression.

We found that the best combination of the following parameters gave us the least RMSE and maximum R-squared value.

Ntree = 101

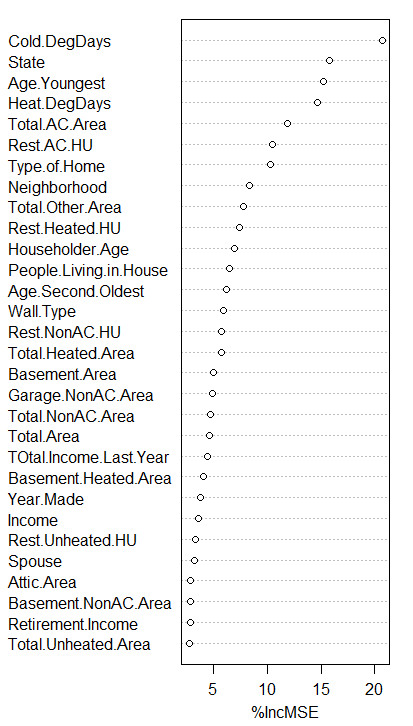
Mtry = 17

Nodesize = 16

Maxnodes = 100

After pruning the tree, we were able to figure out the importance of each variable in the tree. This graph defines the increase in MSE, if that particular variable is removed from consideration. For example, if Cold.Deg Days parameter was removed from the model, than the % increase in MSE would have been approximately 21.73%, i.e. the performance of our model would decrease by 21.73%. Similarly removing Total.Unheated.Area factor from our model will not make a significant impact in the final prediction.

Best parameters passed gives us variable importance as shown :



In the order as described above we decided to analyze top N(starting from 10 to all the features) features which would give the best possible results. After selecting 39 features, there was very slight difference in value of RMSE. So, for similar values of RMSE we picked up the simplest model among all of them. Hence we decided to go ahead with top 39 features using Random Forest Feature Selection.

**2.1.2 Principal Component Analysis**

The primary purpose of using PCA was to reduce dimensionality while retaining most of the variance in the data. In our dataset, we had 17 symbolic features, which would not directly work in PC Analysis. We converted these variables into 84 dummy variables(binary features as per the values in the symbolic features), totalling to 118 total features.

As we know, PCA is a statistical procedure that converts a set of observations of possibly correlated variables into a set of values of [linearly uncorrelated](https://en.wikipedia.org/wiki/Correlation_and_dependence) variables, hence we found a list of all possibly correlated 61 features after comparing their Pearsoncorrelation coefficient in correlation table. All these features were passed into PCA analysis, in order to achieve orthogonal principal components. Now we had 57 actual features and 61 Principal Components for our analysis. We tried various combinations of original features and top principal components(which retained more than 95% of the variance in the data i.e. starting from PC7 till PC61) to find best possible model. All these attempts did not give RMSE value better than what we got from our previous models, as we know it is not necessary that only the top N principal components could give good and accurate results. We then used the method of backward elimination and forward selection, in order to select the best features among these. We finally chose 87 features (a combination of both original features and principal components).

**2.2** **Prediction**

To analyze the historical data and make use of it, in order to derive any future predictions can be termed as Predictive Analysis. We use the electricity consumption data to train and test our models using 10-fold Cross Validation.

**2.2.1 Default Prediction**

This corresponds to the baseline against which we would measure our analysis of other models. The default prediction is the predictor which accounts for no improvement, which could be derived using any independent features. We used the mean value of existing electricity consumption, to predict energy of any other household.

**2.2.2 Linear Regression**

It is the approach for modeling the dependent variable(electricity consumption) against all the other explanatory features. We used linear regression to predict the electricity consumption in three different cases:

i) Using all features

ii) Using features selected from random forest

iii) Features selected after dimension reduction with Principal Component analysis

**2.2.3 Random Forest**

Random decision forest operate by constructing multiple decision trees to overcome decision tree’s habit of overfitting on training data. We used random forest to predict the values of electricity consumption in three different cases:

i) Using all features

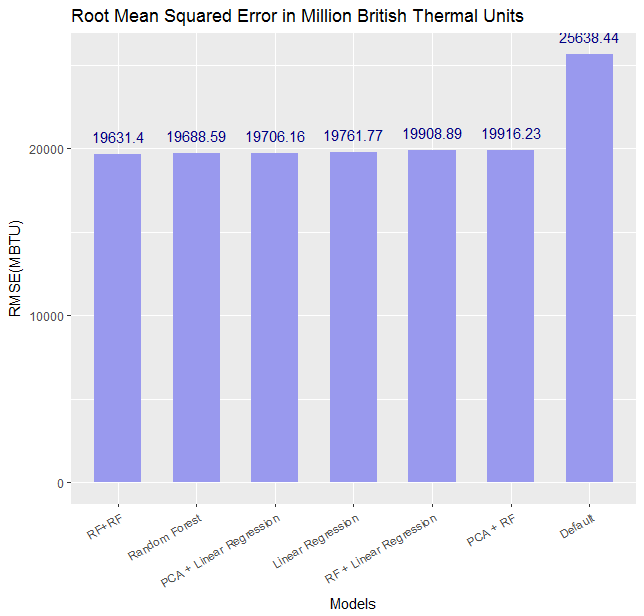
ii) Using features selected from random forest as per the above diagram

iii) Features selected after PC Analysis.

**2.3 Evaluation of prediction**

After analyzing these 7 model on our data we were able to obtain following results.

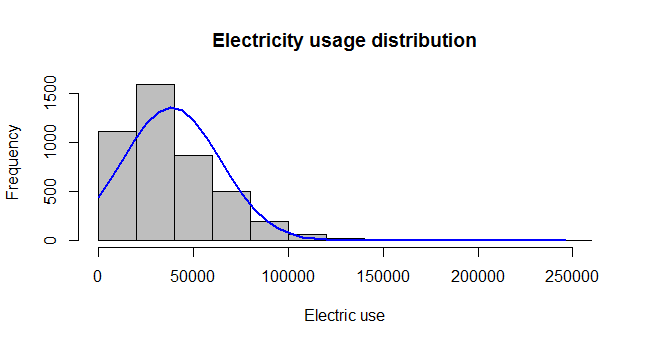
|  |  |  |
| --- | --- | --- |
| **Method** | **R Squared Value** | **Root Mean Squared Error (Million BTU)** |
| All features(51) + Linear Regression | 0.4039074 | 19761.77 |
| All features(51) + Random Forest Regression | 0.41034 | 19688.58824 |
| Random Forest Features (39) + Linear Regression | 0.39824 | 19908.89437 |
| **Random Forest Features (39) + Random Forest Regression** | **0.4120824** | **19631.4** |
| PCA + Backward Selection (87 features) + Linear Regression | 0.4069159 | 19706.16 |
| PCA + Backward Selection (87 features) + Random Forest Regression | 0.395088 | 19916.23 |
| All features(51) + Default Prediction | -0.001800609 | 25638.44 |



Using the graph above, we are able to find the best combination for our model. Features selected with Random Forest (Top 39 features), and using Random Forest regression to predict the final values gave us the best result. If we carefully notice, when we ran random forest over all the features, we got a comparable result, but it would also have been more computationally expensive. Hence according to our analysis, Random Forest variable selection and for prediction gives possibly the best results.

**3. Classification Analysis**

**3.1 Data preprocessing**

Before continuing with our classification task, we need to first create a new binary variable called “Ele.use.level” to get the value for electricity usage level out of the original “BTUEL” column. We took a look at the distribution of the electricity consumptions(Figure \*), and it is right skewed.

We then excluded the observations with an electric consumption value over 100,000(there are 110 such rows). We got the median of BTUEL for the remaining data which is 38650 and this number will serve as our threshold to determine “high”-”low” annual electricity usage. When a household had an electricity consumption higher than 38650 ( in thousands of BTU), their usage level would be defined as “high”, and the usage level would be considered “low” if the consumption value exceeded the threshold. The 50 input variables were still kept as before, but this time the output variable is a new symbolic variable with values of either “low” or “high” instead of BTUEL.

**3.2 Feature Selection using Random Forest**

**3.2.1 Tune parameters for Random Forest**

There are two parameters we tuned for the random forest model, one is “ntree”(number of trees), another is “nodesize” (minimum size of terminal nodes trees in the forest can have).

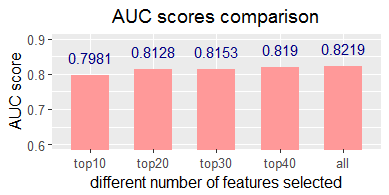
First: We tuned the parameter of “ntree” by running the random forest models many times using different values of the ntree and the values are from three ranges: range 1 to 200 in steps of 10, range 201 to 300 in steps of 20, range 301 to 500 in steps of 30. We applied 10-fold cross validation on these models and used AUC score as an evaluation metric.

We got 51(AUC score = 0.8219) as the best number for ntree, and we used that number for the later analysis. Though we did not notice that some higher number of trees around 400 trees did gives us slightly better AUC score (of the order 0.1 more ) we decided to go ahead with 51 features as this difference was not significantly large and it would reduce the computational costs.

We also tuned the “nodesize” parameter from 2 to 20 with the ntree = 51, and with the same evaluation metric, we chose the nodesize of 13 which gave the best AUC score of 0.823.

**3.2.2 Feature Selection analysis**

We then used this combination of parameters (nodesize=13, ntree=51) in our random forest model to do the feature selection. From the random forest result, we can get a ranking of the feature importance to the model in terms of the decrease of accuracy if excluding that feature. We ran the random forest again four times retaining different number of top features: Top10,Top 20,Top30,Top 40 using the AUC score as the metric. We can see from the plot below that the original model with all features kept gave us the best results. So we kept all the original features in subsequent analysis.



**3.3 Choosing the best Classifier**

**3.3.1 Model choice**

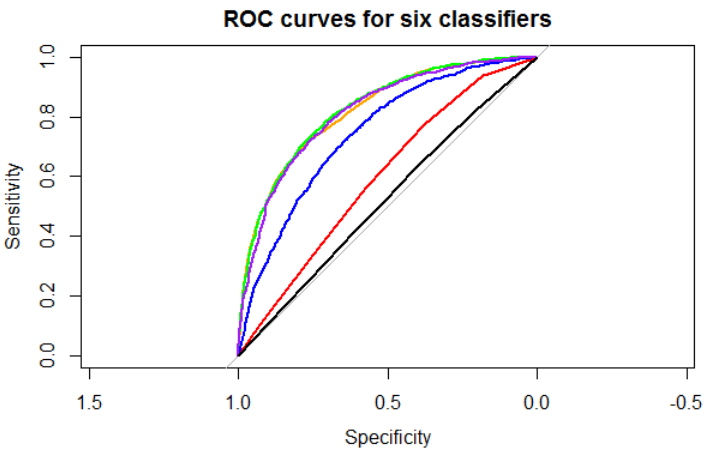
We define “High level electricity user” as positive cases because we want to focus more on those people and define “low level electricity user” as negative cases.

We ran the five classifiers: K Nearest Neighbour, SVM,Naive bayes,Random Forest(ntree=51,nodesize=13) and Logistic Regression.

**3.3.2 Analysis**

**3.3.2.1 Evaluation metric 1: ROC**

We drew the ROC which are shown below(K-NN is red,naive bayes is blue, SVM is green, logistic regression is purple , randomForest is orange.

.

From the above ROC curves and the below AUC scores table, we notice that svm model gave the highest AUC score which means that svm has the highest overall performance among the other classifiers.



**3.3.2.2 Evaluation metric 2: True positive rate (TPR)**

True positive rate would matter more in our scenario since we want to focus more on high energy consumers. We choose 0.5 as the cutoff and we can see clearly from the below table that SVM has a relatively higher true positive rate than other models.



**3.3.2.3 Evaluation metric 3: False positive rate (FPR)**

We also calculated the false positive rate for the five classifiers, since we expected our model to have low “false alarm” rates .The lowest FPR was given by KNN which however also gave the low TPR. The second best FPR score was given by svm model better than the other three classifiers.



**3.4 Results**

1. After doing the feature selection analysis, the original 50 features are still kept in our final classification model since they did gave better results for prediction than other combinations of fewer features.
2. Based on the AUC score, the true positive rate and the false positive rate, support vector machine gives the best performance and is chosen as the model for discriminating between high electricity level user and low level user.

**4. Application of the analysis results to the real world.**

**4.1 Using the regression model**

To encourage household to save electricity, one recommendation for the government is to implement a tiered pricing system for household electricity. From our dataset, we can get the first quantile of the annual electrical use which is 19380(in thousands of BTU) as well as the third quantile which is 49630(n thousands of BTU). Based on these two numbers, we can have three tiers for the electricity usage. First tier is below 19770, second tier is from 19770 to 49630, and the third tier is higher than 49630. For different tiers, set the different price slabs for each tier on per unit of electricity. Since prices vary from year to year, and from state to state, we will just take Pennsylvania as an example here.

The average electricity price of in Pennsylvania in 2016 is 13.76 cents per KilowattHour[[1]](#footnote-0). So the state government of Pennsylvania can set first tier unit price at 13.76 cents, and 14.76 cents for the second tier, 16.76 cents for the third tier. The difference between the third tier price and the second tier price is higher than that between the second tier price and the first tier price, since we want the third tier to control their energy consumption. Some software company can be hired to develop an electricity usage application for consumers to download. The application, similar to Google PowerMeter, helps keep track of the household electricity usage, and gives out warnings to the consumers when their consumption is about to exceed the current tier ceiling value and reminds the consumers that they will have to pay more on the bill if they keep using electricity like before.

Since the electricity is different year by year, and there will always be new families moving to the state, the government can use the model and input variables data to predict the household electricity. Based on that, the government can update the three tiers for electricity consumption, and reset the unit price for the renewed three tiers.

**4.2 Using the classification model**

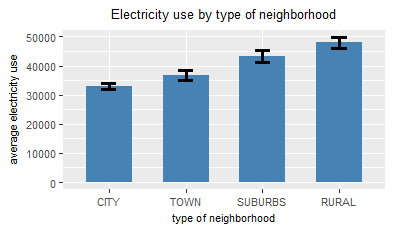
The government can use the household and housing unit profiles to discern whether the household is high electricity level user or low electricity level. Based on this, the government can do the following two things: 1) publish pamphlets promoting the importance of reducing electricity usage for the sake of saving the planet and providing the list of things of the household can follow to save the electricity consumption. Send out the pamphlets to the households with high level electricity usage. 2) impose an electricity usage taxes on these households identified as high electricity users.

**4.3 Specific Features**

Although none of the original features are excluded after analysis, we still took a look into some features which can provide some meaningful insights to help reduce the electric use. The features considered are type of neighborhood with an arity of 4, type of home with an arity of 5, and the type of wall materials with an arity of 9. We drew three bar charts based on the average electricity use of various types of a particular variable as well as the 95% confidence interval error bars. We only went on with the first two features, since the error bars of the wall material variable are so large that the lower bounds of some bars with higher mean value are smaller than the lower mean value of other bars.

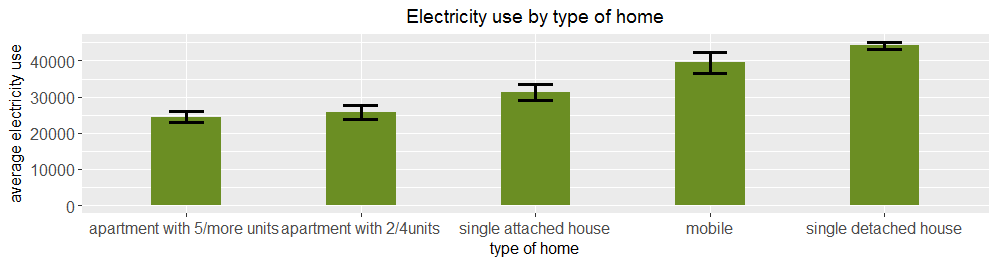
**4.3.1 Type of neighborhood**

We can see from the below bar chart that the houses in the rural area consume more electricity than other type of neighborhood on the average level. The reasons may be that the houses in the rural area have larger space thus consuming more electricity. For example, the houses in the rural area tend to have bigger lawn and the mowers will use much electricity. Another possible factor is that in the rural area, the houses may still use old-fashioned electrical appliances which do not do a good job on energy efficiency. The government can initiate a program encouraging the householders to update their appliances’ energy efficiency by granting some subsidies to replace the old appliances.



**4.3.2 Type of home**

From the bar chart below, we can see that the detached house or mobile house averagely consume more electricity than attached house and apartments with multiple units. This conforms to the reality in that the “population density” is higher in the later types of houses so that residents often share the use of electrical appliances more such as lights and ACs thus reducing energy. The department of urban planning should take this into consideration when they review the real estate project applications, for example, control the number of single detached houses to be constructed. For the NGO and community in college, promoting energy saving at rental companies so that these companies would encourage the tenants to rent the type of homes which can save more energy.



**5. Future Plan**

As we have only data from 1995, future plan should firstly update the data in recent years which can give a new report. And then we can improve our prediction method by using other models and combinations. What’s more, based on the process we have designed, we can analyse other energy consumption and add some features such as every footage consumption or every person’s consumption which can give us new target.

In order to make our result maximize the possibility to succeed in reducing the energy consumption, we have already listed some of the feasible ways to implement in ‘Application of the analysis results to the real world’ part.

**6. Risks**

i) Our evaluation for prediction was very closely tied between using 39 features selected from Random Forest or go with all the features. On the basis of current data, we ruled out the possibility of using all features in order to achieve more complexity, but it is possible that if we run this model on larger set of data, all features perform better.

ii) While classification, there was a risk that we may not end up with the best possible result if we choose wrong model for classification, considering the fact that all the three evaluation metrics give very close results. It is possible that our result may change when we run this model on a different dataset.

iii) Our understanding of changeable features, might not be valid for each scenario and hence it is subject to interpretation.

**Statement of Contribution**

--Abhishek Jain

Data Cleansing

PCA Analysis

Naive Bayes Classification

KNN Classification

Random Forest Modeling in Regression

Wrote Prediction part and Exec Summary components in the final report

Team Coordination

--Bo Wang

Correlation and PCA analysis

SVM Classification

Linear regression and random forest modeling

Application and business analysis

Report writing

--Rashmi Upadhyaya

Data Cleansing

Random Forest for Regression and tuning

Final Report

Final presentation slides and video

--Yuyue Jiang

Prepared codes for tuning parameters of random forest models in regression and classification.

Visualized analysis results in plots.

Wrote about data preprocessing, classification analysis and real-world application of analysis results in the final report.

1. Source: https://www.eia.gov/electricity/monthly/epm\_table\_grapher.cfm?t=epmt\_5\_6\_a [↑](#footnote-ref-0)